MSI Journal of
Economics and Business Management (MSIJEBM)

Analysis of Large Language Model in Creating User Personas: A Comparative Study Across Cultures

Atefeh Kasiri[1*]

[1*] Germany.

*** Correspondence:** Atefeh Kasiri

The authors declare that no funding was received for this work.

Received: 30-September-2025
Accepted: 10-October-2025
Published: 21-October-2025

Copyright © 2025, Authors retain copyright. Licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. https://creativecommons.org/licenses/by/4.0/ (CC BY 4.0 deed)

This article is published in the **MSI Journal of Economics and Business Management (MSIJEBM)**

ISSN 3049-141X (Online).

The journal is managed and published by MSI Publishers.

Volume: 2, Issue: 10 (October-2025)

ABSTRACT: The emergence of large language models (LLMs) has changed the world and human lives in many different ways. Nowadays there is a new idea that LLMs are a big database of people's opinions, therefore, It is possible to replace them with humans in some parts of the design process, while LLMs may not capture all the individual and cultural differences, and it can lead to discrimination against a group of people.

This thesis investigates the representativeness of LLM-generated user personas by comparing three different persona creation approaches—LLM-solely, LLM-auto, and LLM-summarizing—across two cultural contexts: Germany and Iran. By collecting survey data from real participants in both countries, this study establishes a ground truth for evaluating the generated personas with LLM.

Introduction

User persona is considered an essential component in the Human-Centered design (HCD), which is a hypothetical user archetype that models real user behaviours, complementing direct user participation in the software engineering process [1], embodying their unique traits, behaviours, and requirements. Old ways of user persona creation necessitate frequent interaction between development teams

and the end users to specify “what users want” from “what software engineers think users want” and this old way of user persona creation with constant human interaction has its challenges. and old ways of persona creation has still the problem of finding representative group of users and interacting with them so there is the new idea of leveraging LLMs for persona creation [2].In response to these challenges, researchers and practitioners have begun exploring the potential of large language models (LLMs), such as OpenAI’s GPT series, to automate parts of the persona creation process. These models, trained on vast datasets containing human language, have been proposed as proxies for human perspectives—able to generate persona profiles through well-crafted prompts. This approach promises efficiency and scalability, but also raises critical concerns [2] Previous studies suggest that while LLMs can produce coherent and seemingly plausible personas, they often fail to capture nuanced socio-demographic differences, and may encode or reproduce harmful biases when tasked with adopting human roles or viewpoints. This is especially problematic in HCD, where accurate and inclusive user representation is essential for ethical and effective design [3].

The objectives of this work is to investigate the following hypothesis

Null Hypothesis 1(H0): There is no significant difference in the representativeness1 of user personas generated by Language Models (LLMs) for individuals with diverse cultural backgrounds.

Null Hypothesis 2 (H0): Nationality (Iran/Germany) and approach (LLM-auto / LLM-Summarizing) interaction has no significant effect on the representativeness of the personas.

RELATED WORK

This part aims to give insights about LLMs and user personas and the effects that LLMs had on the persona creation

a. Large Language Models (LLMs)

Large Language Models (LLMs) have transformed the fields of natural language understanding (NLU) and generation. They have deep language comprehension, human-like text generation ability and problem-solving skills, which make them indispensable

across a wide range of domains [4]. Despite their strengths, LLMs present ethical and practical challenges. Issues such as content moderation, filtering, and accountability remain unresolved. These models can generate misinformation, hate speech, and biased content, raising serious ethical concerns [5] There is work on the creation of generative agents, an interactive sandbox environment like The game Sims, allowing end users to interact with a small town of twenty-five agents using natural language, To enable generative agents, they outline an agent architecture that stores, synthesizes, and applies relevant memories to produce believable behaviour using a large language model

[6]. Generative agents can have different applications for example, social simulacra have shown the ability to create stateless personas that generate conversation threads in online forums for social prototyping. Such agents have applications in social prototyping, online forums, virtual reality environments, and robotics. When paired with multimodal systems, they enable advanced human behavior simulations to test social systems and design interactions [7]. LLMs primarily focus on generating responses rather than ensuring those responses are correct. OpenAI has acknowledged that their latest model, GPT-4, released in March 2023, "is not fully reliable (it hallucinates facts and makes reasoning errors)" and advises caution when using GPT-4's outputs, especially in contexts where reliability is crucial [8]. As survey research works with words in the questions and answers so it is expected that LLMs affect this domain because different tasks in survey research can be done with LLMs [9]. CloudResearch's Chief Technology Officer, Jonathan Robinson, explained ''Our team has been working on this advancement for years. Survey researchers kept telling us about problems they were having with attention and data quality. It's also always been difficult to find people from hard-to-reach groups. So, they thought, 'What if we just got rid of the people altogether? That would solve a lot of problems" [10] Employing LLMs could greatly enhance the efficiency of these tasks. Potential applications of LLMs in survey research include simulating human responses, predicting public opinion, creating new survey questions with generative AI, filling in missing data, providing feedback to respondents, and reporting survey responses as interaction data. But in this case, the result may contain bias, stereotypes, discrimination or extraordinary norms that affect the survey result or there can be a problem of contextual understanding, and it can cause non-relevant responses to the survey questions (hallucination), Using LLMs in

survey research also raises ethical concerns, particularly regarding the replacement of human participants with AI-generated responses. Questions arise about the appropriateness and acceptance of LLMs as respondents in psychology research within the research community. Additionally, there are concerns about whether LLMs can accurately mimic human cognition and emotions when responding to surveys focused on psychology and behaviour-related topics [9]. Regarding the replacement of human participants with LLMs usually survey distributors carefully select representative samples of the human population to ensure valid results and to account for demographic differences. For LLMs to serve as effective replacements, they must be able to capture the impact of social identities like gender and race, there are two fundamental limitations in how current LLMs are trained that hinder this capability mis portrayal and group flattening [11].

B. User Persona in Human-Centered Design (HCD)

Requirements Engineering (RE) seeks to identify the goals, needs, and expectations of end users in the software development process [12]. Traditional RE requires ongoing interaction with users to ensure that the system being developed reflects their true needs rather than developers' assumptions. However, maintaining regular access to users throughout the software development process can be challenging due to several factors: finding a representative set of users to collaborate with the development team; having access to only a limited subset of users, which may not fully represent the diversity of end-user perspectives; ensuring that interactions with various end users are effective; and obtaining timely access to these users to avoid delays in the design and development of the software. To address this, Alan Cooper introduced the concept of the *persona* in Human-Computer Interaction (HCI) a fictional but data-informed character that captures the behaviors, goals, and challenges of a user group [13]. A persona represents a composite profile of target users who exhibit similar behavioral characteristics [14]. Personas are fictional, but detailed and specific characters that represent various user types. They facilitate a shared understanding of users' demographics, needs, behaviours, motivations, and challenges [12]. In the process of software engineering, it specifies "what users want" from "what software engineers think users want" and during the requirement engineering

(RE) process, the more real users participate the better requirement specifications satisfy the end users [13].

However, using personas can cause certain risks, such as inaccurately representing the user groups. To mitigate the limitations of traditional personas, the use of data-driven methods for generating personas is suggested, which can support business-to-business software development companies [12]. Recently a new persona tool has been introduced called Persona Gen, this tool utilizes the GPT-4 model and knowledge graphs to generate personas from well-processed user feedback. PersonaGen boasts three main features:

1. It is the first tool to employ the GPT-4 model for cleaning, integrating, predicting, and analyzing user feedback.

2. It builds knowledge graphs to organize data attributes.

3. It classifies persona traits and generates recommendations [15]. The term "human-centered AI" is becoming more prevalent, emphasizing the idea that AI should serve people and addressing rising concerns about AI's potential to exploit and mislead. However, "human-centered AI" holds different meanings for different individuals [16]. Human-centered AI (HCAI) involves creating artificial intelligence technologies that focus on human needs, values, and capabilities at their core. This approach ensures that AI systems are designed to enhance human abilities and well-being rather than replace or diminish human roles. It considers the ethical, social, and cultural implications of AI, ensuring these systems are accessible, usable, and beneficial to all segments of society. HCAI is closely related to Human-AI interaction, a field that studies how AI and humans communicate and work together. In Human-Centered AI, designers and developers work collaboratively across disciplines, often including psychologists, ethicists, and domain experts, to develop AI that is transparent, explainable, and accountable. This approach aligns with the broader ethical AI movement and stresses the importance of creating AI systems that uphold human rights, fairness, and diversity [17]. Human-centered design (HCD) focuses on placing the human at the core of interactive system design but there is a new idea that can be achieved without actively involving the human user in the process? If we consider LLMs as a big database of people's opinions then it is possible to replace them with humans in some parts

of the design process, it is reasonable to use LLMs if the information we get from them is similar to the information we get from human but if not we should not use LLMs as shortcuts in the HCD process meanwhile LLMs have their challenges like biases, prompting and system specification also as time passes different LLMs are generating more content and they are interacting with each other so the data pool that is their feed is increasingly filled with LLMs generated data and human's experience role is becoming weaker so from some point LLMs start reflecting their perspective [6].

c. LLM-Created User Persona

in the field of AI and Human-AI collaboration it is supposed that human and AI systems can work with each other to achieve goals [18] large language models (LLMs) can effortlessly generate textual outputs from prompts, leading to suggestions for using tools like ChatGPT to create user personas. Resources such as online videos and blog posts demonstrate how to generate a persona with ChatGPT using prompts like: "Generate a user persona about a busy mum who lives in a city and wants to ensure her kids eat healthily without compromising her time for hobbies and career." [1] Prompts like these can be entered into the web chat to create a user persona with believable details [18], However, personas generated by LLMs without any empirical data are entirely fictional and lack the realistic component implied by the definition of a persona as 'fictional, yet realistic' [18].

LLMs are trained on massive datasets from sources like Wikipedia, BookCorpus, and web crawls [19], LLMs often adopt stereotypes, misrepresentations, derogatory and exclusionary language, and other harmful behaviors that disproportionately impact already vulnerable and marginalized communities, these are all types of social biases and stereotypes [20]. Social bias in large language models (LLMs) is worrisome because it can harm targeted social groups. These harms can be representational, either portraying certain groups negatively or failing to represent them at all, or allocational, denying specific groups access to opportunities or resources [21].

LLMs have the ability to generate human-like text and adapt to various natural language processing (NLP) tasks, the impressive capabilities of these models have revolutionized language model development. [1]https://www.youtube.com/watch?v=watkVh1U_ko.

Instead of training task-specific models on relatively small datasets, researchers and practitioners can now use LLMs as foundational models that can be fine-tuned for specific purposes [20]. Even without fine-tuning, these foundational models increasingly support few- or zero-shot capabilities for a wide range of scenarios, such as classification, question-answering, logical reasoning, fact retrieval, and information extraction [20] [22]. LLMs can have their own “personalities” that can be evoked in-context. While LLMs often behave like the average person, their personality profiles can be modified, such as by adjusting the context to be more or less emotional [36]. LLMs can even exhibit "personalities" based on in-context prompting. Their tone, behavior, and even reasoning style can be adjusted, for example, by instructing them to "act as an elementary school teacher" or "behave like a scientist.", These "persona-assigned LLMs" not only enable engaging and enjoyable interactions through personalization but also have numerous practical applications due to their ability to mimic human behavior. there is a study that examines 19 diverse personas that represent a wide range of sociodemographic factors, such as race, religion, and political affiliation. The result reveals that socio-demographic personas not only affect the reasoning ability of LLMs but also expose deeply ingrained stereotypical biases within them [8].

Even shows when comparing personas within a single socio-demographic group (e.g., religion), model biases lead to varying performance levels. For example, the Jewish persona performs better on Science, technology, engineering, and mathematics (STEM) datasets, the Atheist persona outperforms Christians in the Sciences, and the Obama Supporter persona surpasses the Trump Supporter persona in ethics [8]. It is undeniable that the existence of such biases in a system can make it inapplicable in real-world applications because it can cause disastrous consequences [19].

D. human and LLM-generated user persona

There was an online statement which was game-changing for the user experience (UX) community, entitled “Can ChatGPT Replace UX Researchers? An Empirical Analysis of Comment Classifications” The post was mentioned by Jakob Nielsen, researcher in HCI and commented “This doesn’t mean that ChatGPT can analyze user *behavior*” [1]. only it is likely a time-saver in grouping non-behavioral questionnaire responses. (ChatGPT is

known to be great at summarizing and classifying text without understanding what it means.)" [23]. [1]https://www.linkedin.com/posts/jakobnielsenphd_can-chatgpt-replace-ux-researchers-an-empirical-activity-7067298800259760128-ZmsS/

This naturally brings us to consider the role of large language models (LLMs) in creating user persona. This is a potential application for UX researchers using these advanced technologies. Personas are crucial in UX research, serving as fictional yet realistic representations of users. They are considered fictional because they are narratives aimed at capturing specific characteristics of target user groups. They are realistic because they are based on data collected from actual users [23].

METHODOLOGY

As mentioned earlier, Requirements Engineering (RE) tasks often depend on direct engagement with stakeholders to understand their goals and expectations. However, this approach is resource-intensive and sometimes impractical, especially when repeated user access is needed. To address this, personas are commonly used as efficient stand-ins for real users. Personas are typically created using data from methods such as contextual inquiry, focus groups, or surveys [24]. Since these methods are the main way to collect information, creating personas can be time-consuming and require a lot of resources. To solve this, the automated creation of personas through data science methods has emerged, known as "Data-Driven Persona Development (DDPD)." Salminen et al. provide a comprehensive overview of DDPD research [25]. A major obstacle to the mentioned methods is the significant effort needed to produce high-quality representations of the gathered data. To address this, researchers explore human-AI workflows where Large Language Models (LLMs) assist in various tasks related to user data analysis, However, LLMs also have technical limitations that can affect the quality of the analysis, such as biases and the generation of inaccurate information due to biased training data and limited domain knowledge [26].

This work is an extension of a paper on qualitative user data (synthetic survey responses and survey responses from real people), and it shows that the most accurate and empathy-driven personas emerge when human experts lead the process of defining key user groups

while leveraging LLMs for summarizing data. You can find the source code of the project in GitHub repository for this work.[1]

Research Design

What I want to do in this work is "Analysis of Larg Language Models in Creating User Personas: A Comparative Study Across Cultures" I chose 2 different countries for this study Germany and Iran and I prepared 30 survey-response set as user data from each of the mentioned countries and I consider the user data as the ground-truth data because eventually the goal of persona is to fulfil the user need so the better is the persona which is closer to user data. [1]https://github.com/atefehkasiri/cross-culture-persona-generation

I create the persona in 3 different ways

- First, I prompt the LLM to fill out the form 2 times each time as a user from the mentioned countries and then I ask the LLM to create the personas
  according to the forms and it gives 2 personas for the mentioned countries.
- Second, I give the user data (survey responses) to the LLM and ask the LLM to create the persona based on the user data and the
- The third way is I identify Key characteristics in the surveys and group them and ask the LLM to create persona based on the groups.

Figure 1 demonstrate Persona creation workflows

I used statistical analysis and compared the personas created in 3 ways with the real user data (ground truth). *Study Population and Questionnaire*

The study population comprised individuals from Iran and Germany, The selection of participants was based on a convenience sampling method, a non-probability sampling technique that allows for the selection of individuals who are readily available and willing to participate, for this work a total of 60 respondents were included in the study, with 30 participants from Iran and 30 from Germany. This study is conducted in 2 different countries Germany and Iran, Based on the amount of data that is created in these countries, the second country with the most amount of data centers is Germany also I chose Iran because it is not even in the first 15 countries [27]. The other reason for these countries is

the number of internet users in comparison with the population, which in Germany is 93.3% and in Iran is 88.8% of the population of the countries [28].

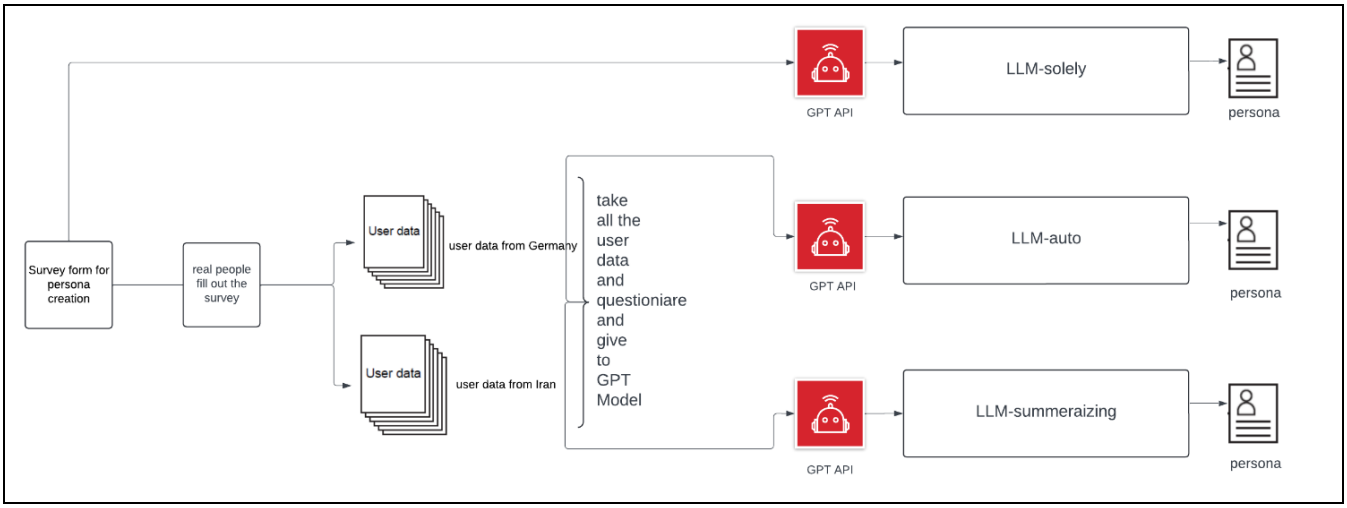

Figure 1: Persona creation workflows for LLM-Solely /auto /summarizing.

As social media has grown in popularity and become a key part of the global economy, understanding the factors that shape consumer attitudes toward social media platforms has become a primary focus for researchers in both academia and industry [29] This study is focused on creating a user persona in the field of Social Media, Communication, and Technology Adoption; this user persona can have many different applications to improve product design, user experience, marketing strategies, and user research methodologies.The structure of the questionnaire is designed by me in a way to collect comprehensive data that contributes to developing a user persona which can differentiate the cultural differences. Each section is included to capture key aspects of user behavior, preferences, and motivations in the field of Social Media, Communication, and Technology Adoption.

Prompt Engineering for LLM-based Persona Creation

Prompt engineering plays a crucial role in the performance of LLMs. A prompt serves

as a set of instructions given to an LLM, allowing users to tailor its responses, optimize its functionality, and refine its capabilities [30]. The General structure of the prompts I used for the work has three parts, as shown in Table 1. A well-structured prompt consists of four main parts: instructions, context, input data and output indicators [31].

The prompts that I used for this paper cover these 4 parts, First I give the questionnaire in .txt format and the user data in the .csv format to the model and then clarify the task that the model should do and the rules that should follow and what to generate as output. In the first approach for persona generation in this work, which is LLM-Solely no user data is given to the model. This means the model has to rely entirely on its pre-trained knowledge and it was a zero-shot prompting but for the LLM-auto and LLM-summarizing approaches the user data was given to the model so it is a few-shot prompt engineering [32].

Prompt
Here are the survey questions
Here is the user data
The task that the model should do
The rules that the model should follow while doing the task

Table 1 Prompt design. The input data(yellow-highlighted), persona generation task(blue-highlighted),rules LLM should follow(green-highlighted)

LLM-Solely: In the first workflow within a single prompt, I gave all the survey questions and instructed it to answer all the questions as a person from different countries to create a user persona.

LLM-Auto: In this approach, I provide all user data (survey results) within a single prompt and ask LLMs to generate a minimum number of personas based on the given data, allowing LLMs to independently identify traits in the data without being influenced by human biases.

LLM- Summarizing: in the final workflow, I first clustered user data based on key traits I identified and the K-means clustering algorithm [33] and I created a minimum number of data groups. Then, by providing one group of user data per prompt, I instruct LLMs to generate personas individually.

Data Collection

LLMs like GPT are non-deterministic, meaning their outputs can vary across sessions—even when given the same input [34]. While working with LLM we should keep in mind

that Chat completions are non-deterministic, meaning the model's outputs can vary between requests [35]. although for using the OpenAI API different parameters can affect the non-determinism like “Temperature” which controls the randomness of next words during text generation, and the other parameter is “Top-p” also known as nucleus sampling, which is a parameter that determines the number of possible words the model considers. table 2 shows GPT’s answer regarding its determinism, that it is saying the same answer to one prompt but in different ways.

model="gpt-4o-mini",	messages=[{"role": "system", "content": "You are a helpful assistant."}, { "role": "user", "content": "do you answer same to a same question when I ask it many times? in other words are you deterministic in what you say? " }
First answer	"I strive to provide consistent answers to the same question, but there can be slight variations in wording or detail depending on the context of the conversation and how the question is framed each time. While I'm designed to be consistent, the randomness of language generation means that repeated questions may yield similar but not identical responses. If you're looking for a precise answer, feel free to ask your question again, and I'll do my best to respond consistently! "
Second answer	"I strive to provide consistent and accurate answers to the same questions, but there may be slight variations in phrasing or detail depending on the context and how the question is presented. While I'm designed to be consistent, the complexity and nuance of language can lead to different responses at times. However, the core information and intent of the response will generally remain the same. If you have a specific question, feel free to ask, and I'll do my best to provide a clear and relevant answer! "

Table 2 GPT answers regarding determinism

ANALYSIS AND RESULTS

As the purpose of this work is to compare user persona creation across cultures and I chose 2 countries for my work I have 2 sets of population samples with relatively similar demographic features, these are relatively similar because the amount of salary is dependent on the countries. The nominal income levels in each country based on their local currencies (Iranian Rial and Euro) are different.

To measure representativeness, a custom metric called the similarity score was introduced. This score indicates the percentage of similarity between the options that each individual chose and the options that the LLM-generated persona chose in the survey, in a fixed 86-

option survey. For example, if a participant selected 43 of the same options as the LLM-generated persona, their similarity score would be 50%.

Analysis of LLM-Generated Personas across Cultures for LLM-Solely Approach

In this section I analyze the personas that LLM created in the LLM-Solely approach with the ground truth data. In this approach no user data is given to the LLM, only the survey questions are given to the model and prompted to create a persona from Iran and Germany and answer the survey questions based on the personas.

I compared the generated persona with each person in my sample to see how representative the LLM-solely-generated persona of the real people (ground truth data) is. Table 3 shows how representative the LLM-solely-generated persona is to the Iranian sample, the maximum similarity was 57.5% in one case and the minimum similarity was 37.5% and on average it was 46.25% representative of the Iranian people. Table 4 shows how representative LLM-solely-generated persona is to the German sample, the maximum similarity was 60% in one case and the minimum similarity was 38.75% and on average it was 50.79% representative of the Iranian people.

	Min	Max	Mean
Percentage of similarity between LLM-Solely-Generated persona and individuals in the Iranian sample	37.5%	57.5%	46.25%

Table 3 Similarity score between LLM-solely generated persona and Iranian sample

	Min	Max	Mean
Percentage of similarity between LLM-Solely-Generated persona and individuals in the German sample	38.75%	60%	50.79%

Table 4 Similarity score between LLM-solely generated persona and Iranian sample

Considering the tables above, I can conclude that the LLM-solely generated persona for Germans was more representative of individuals in the German sample in comparison with the Iran sample.

To test my hypothesis, I did an independent samples t-test on the 2 samples based on the variable similarity score. The mean similarity score for the German persona is 50.7 that is more than the mean similarity score for the Iran persona 46.2. As sig in the Leven's test for equality of variances is (.808 > .05) there is no significant difference between the variances of the 2 samples based on the sig.(2-tailed) which is considered as p-value and it is smaller than 0.5. than I can reject the null hypothesis which is "There is no significant difference in the representativeness1 of user personas generated by Language Models (LLMs) for individuals with diverse cultural backgrounds." I can present that there is a significant difference between the representativeness of LLM-Generated personas for these 2 cultures.

Analysis of LLM-Generated Personas for Iranians for second and third approach

The second approach is LLM-auto, I gave the Iran sample and questionnaire to LLM and asked to generate a persona based on the given data, it generated 2 personas that I call them LLM-auto-a and LLM-auto-b, I compared each of them with each individual in Iran sample(ground-truth data) to see how representative is each persona of the sample, for LLM-auto-a the max similarity with one individual was 61.25% and min was 37.5% and on average 50.45% and respectively for LLM-auto-b the numbers are 61.25, 36.25,48.29 percent.

The third approach is LLM-summarizing which is the process includes human manipulation, I clustered the people data in 2 clusters and I asked LLM to generate a persona based on each cluster and I called them LLM-summarizing-a and LLM-summarizing-b and I did the same analysis for each persona, similarity score for LLM-summarizing-a and persona was max 65% representative of an individual min was 46.25 and the average was 53.39 and the same process with second cluster and LLM-summarizing-b persona and as follows 61.25, 37.25 and 51.56 were the percentages of representativeness. In the table 5 below you can see all the numbers that are in percent.

Analysis of LLM-Generated Personas for Germans for the second and third approach

I compared the LLM-auto-generated persona with all the people in the German sample to see how representative it is of individuals and the min was 37.75% and max was 63.75% and on average it was 54.87% of similarity in LLM-auto -generated persona and the individuals.

And for the LLM-summarizing, I clustered the German sample to 2 clusters and prompted LLM to create a persona for each cluster so it created 2 personas that LLM-summarizing-a for the first cluster that the min similarity was 43.75 and the max was 58.75 and on average 54.37 percent of similarity between LLM-summarizing-a persona and people on the first cluster, and respectively for LLM-summarizing-b persona and individual in second cluster was 36.25, 58.75, and 53.03 percent.

		LLM-solely	LLM-auto-a	LLM-auto-b	LLM-summarizing-a	LLM-summarizing-b
Percentage of similarity between LLM-Generated persona and individuals in the Iranian sample	Min	37.5	37.5	36.25	46.25	37.5
	Max	57.5	61.25	61.25	65	61.25
	Mean	46.25	50.45	48.29	53.39	51.56

Table 5 Min, Max, Mean of similarity scores in 3 different approaches for the Iranian sample

		LLM-solely	LLM-auto	LLM-summarizing-a	LLM-summarizing-b
Percentage of similarity between LLM-Generated persona and individuals in the German sample	Min	38.75	38.75	43.75	36.25
	Max	60	63.75	58.75	58.75
	Mean	50.79	54.87	54.37	53.03

Table 6 Min, Max, Mean of similarity scores in 3 different approaches for the German sample

For the second approach, LLM-auto when I gave the Germans data and the questionnaire and prompt to create a user persona based on the given data, LLM generated one persona that is while in the LLM-auto it generated 2 personas based on Iranians data, no one can determine a certain reason for the LLM's behaviour, but I thought maybe one possible

reason for this behaviour was that Iranians sample was more diverse than German, to test this comparison I calculated the variance of users answers to each option of the questionnaire which shows how far apart people's answers to the question were and the sum of all of these for the whole survey shows how diverse people's opinions in each sample were. The sum of the variances of options in the Iranian sample was 32.84 and for the German sample, it was 31.13. this shows Iranian sample was a bit more diverse.

Analysis of LLM-auto and LLM-summarising Approaches Across Cultures

In order to estimate how the mean of the similarity score for the user personas will change according to the nationality (Iranian/ German) and approach (LLM-auto / LLM-summarizing) factors I used 2 way ANOVA (Analysis of Variance) statistical test.

Here is the result of the 2 way ANOVA and based on the P-value for each factor and their interaction it is concluded that

[1]H0 is rejected as the P-value is (.001<0.05) , so nationality has a significant effect on the similarity score

[2]H0 failed to be rejected as the P-value is (.304>0.05), so different approaches have no significant effect on the similarity score

[3]H0 is rejected since the P-value is (.026<0.05), therefore Nationality and approach interaction has a significant effect on the similarity score

The bar graph below interprets the same interaction of nationality and approach.

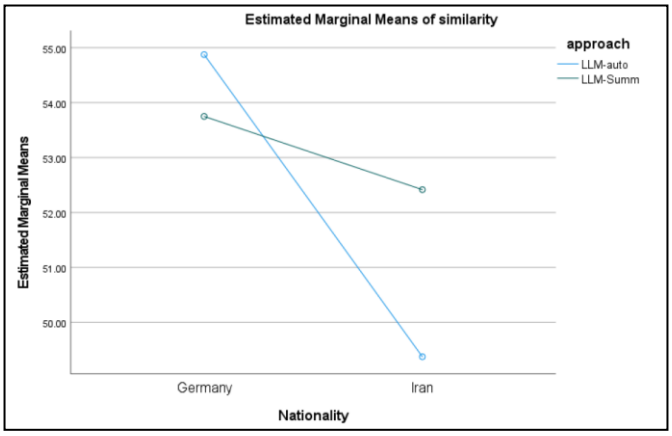

Figure 2: Bar graph of the interaction of nationality and approach

Research Overview

This paper is focused on creating a user persona in 3 different approaches by utilizing LLM across cultures and aims to show how representative the personas of real individuals are. For this purpose, I created user personas in 3 approaches for 2 different nationalities, and then I compared the LLM-generated user personas with the ground truth data that was collected from real individuals to see if there is a difference in the representativeness of user personas across cultures and in between the approaches.

The first part of the research analyses the user personas generated with an LLM-solely approach across cultures and answers to the first research question

• Research Question 1 (RQ1): Can large language Models create user personas representing individuals from different backgrounds?

I approved that the null hypothesis is rejected for the first research question, so there is a significant difference in the representativeness of user personas generated by LLMs in the LLM-solely approach for individuals with diverse cultural backgrounds.

The second part of the work focuses on LLM-auto and LLM-summarizing approaches across cultures and answers the second research question

• Research Question 2 (RQ): How does the relationship between nationality and the user-data-based approach (LLM-auto or LLM-summarizing) for creating personas influence the representativeness of the resulting personas?

It is demonstrated that the null hypothesis for research question 2 is rejected, so Nationality (Iran/Germany) and approach(LLM-auto / LLM-Summarizing) interaction have a significant effect on the representativeness of the personas.

REFERENCES

1. D. Karolita, J. McIntosh, T. Kanij, J. Grundy, and H. O. Obie, “Use of personas in Requirements Engineering: A systematic mapping study,” *Information and Software Technology*, vol. 162, p. 107264, Oct. 2023, doi: 10.1016/j.infsof.2023.107264.

2. A. Schmidt, P. Elagroudy, F. Draxler, F. Kreuter, and R. Welsch, “Simulating the Human in HCD with ChatGPT: Redesigning Interaction esign with AI,” *interactions*, vol. 31, no. 1, pp. 24–31, Jan. 2024, doi: 10.1145/3637436.

3. S. Gupta *et al.*, “Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs,” Jan. 27, 2024, *arXiv*: arXiv:2311.04892. doi: 10.48550/arXiv.2311.04892.

4. Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, “A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly,” *High-Confidence Computing*, vol. 4, no. 2, p. 100211, Jun. 2024, doi: 10.1016/j.hcc.2024.100211.

5. M. A. K. Raiaan *et al.*, “A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges,” *IEEE Access*, vol. 12, pp. 26839–26874, 2024, doi: 10.1109/ACCESS.2024.3365742.

6. J. S. Park, J. C. O’Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, “Generative Agents: Interactive Simulacra of Human Behavior,” Aug. 06, 2023, *arXiv*:arXiv:2304.03442. doi: 10.48550/arXiv.2304.03442.

7. K. Orland, “So what is ‘the metaverse,’ exactly?,” *Ars Technica*, Nov. 07, 2021. https://arstechnica.com/gaming/2021/11/everyone-pitching-the-metaverse-has-a-different-idea-of-what-it-is/

8. M. Scanlon, F. Breitinger, C. Hargreaves, J.-N. Hilgert, and J. Sheppard, “ChatGPT for digital forensic investigation: The good, the bad, and the unknown,” *Forensic Science International: Digital Investigation*, vol. 46, p. 301609, Oct. 2023, doi: 10.1016/j.fsidi.2023.301609.

9. B. J. Jansen, S. Jung, and J. Salminen, “Employing large language models in survey research,” *Natural Language Processing Journal*, vol. 4, p. 100020, Sep. 2023, doi: 10.1016/j.nlp.2023.100020.

10. A. M. PhD, “CloudResearch Revolutionizes Online Survey Research with Virtual Recruitment of AI Participants,” *CloudResearch*, Mar. 31, 2023. https://www.cloudresearch.com/resources/blog/virtual-participant-recruitment/

11. A. Wang, J. Morgenstern, and J. P. Dickerson, “Large language models that replace human participants can harmfully misportray and flatten id ntity groups,” Feb. 03, 2025, *arXiv*: arXiv:2402.01908. doi: 10.48550/arXiv.2402.01908.

12. Y. Wang *et al.*, “Who Uses Personas in Requirements Engineering: The Practitioners’ Perspective,” Mar. 23, 2024, *arXiv*: arXiv:2403.15917. doi: 10.48550/arXiv.2403.15917.

13. D. Karolita, J. McIntosh, T. Kanij, J. Grundy, and H. O. Obie, “Use of personas in Requirements Engineering: A systematic mapping study,” *Information and Software Technology*, vol. 162, p. 107264, Oct. 2023, doi: 10.1016/j.infsof.2023.107264.

14. T. Miaskiewicz and K. A. Kozar, “Personas and user-centered design: How can personas benefit product design processes?,” *Design Studies*, vol. 32, no. 5, pp. 417– 430, Sep. 2011, doi: 10.1016/j.destud.2011.03.003.

15. X. Zhang *et al.*, “PersonaGen: A Tool for Generating Personas from User Feedback,” in *2023 IEEE 31st International Requirements Engineering Conference (RE)*, Hannover, Germany: IEEE, Sep. 2023, pp. 353–354. doi: 10.1109/RE57278.2023.00048.

16. T. Capel and M. Brereton, “What is Human-Centered about Human-Centered AI? A Map of the Research Landscape,” in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg Germany: ACM, Apr. 2023, pp. 1–23. doi: 10.1145/3544548.3580959.

17. “What is Human-Centered AI (HCAI)? — updated 2024,” *The Interaction Design Foundation*, Feb. 11, 2024. https://www.interaction-design.org/literature/topics/human-centered-ai

18. “Creating UX Personas with ChatGPT,” *www.youtube.com*. https://www.youtube.com/watch?v=watkVh1U_ko

19. R. Navigli, S. Conia, and B. Ross, “Biases in Large Language Models: Origins, Inventory, and Discussion,” *J. Data and Information Quality*, vol. 15, no. 2, pp. 1–21, Jun. 2023, doi: 10.1145/3597307.

20. I. O. Gallegos *et al.*, “Bias and Fairness in Large Language Models: A Survey,” Jul. 12, 2024, *arXiv*: arXiv:2309.00770. doi: 10.48550/arXiv.2309.00770.

21. S. Kolisko and C. J. Anderson, “Exploring Social Biases of Large Language Models in a College Artificial Intelligence Course,” *AAAI*, vol. 37, no. 13, pp. 15825–15833, Jun.2023, doi: 10.1609/aaai.v37i13.26879.

22. P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing,”Jul. 28, 2021, *arXiv*: arXiv:2107.13586. doi: 10.48550/arXiv.2107.13586.

23. De Paoli, S. (2023) Improved prompting and process for writing user personas with llms, using qualitative interviews: Capturing behaviour and personality traits of users, arXiv.org. Available at: https://arxiv.org/abs/2310.06391

24. W. C. Tomlin, *UX Optimization: Combining Behavioral UX and Usability Testing D ta to Optimize Websites*. Berkeley, CA: Apress, 2018. doi: 10.1007/978-1-4842-3867-7.

25. J. Salminen, K. Guan, S.-G. Jung, and B. J. Jansen, “A Survey of 15 Years of Data- Driven Persona Development,” *International Journal of Human–Computer Interaction*, vol. 37, no. 18, pp. 1–24, Apr. 2021, doi: https://doi.org/10.1080/10447318.2021.1908670

26. J. Shin, M. A. Hedderich, B. J. Rey, A. Lucero, and A. Oulasvirta, “Understanding Human-AI Workflows for Generating Personas,” in *Designing Interactive Systems Conference*, IT University of Copenhagen Denmark: ACM, Jul. 2024, pp. 757–781. doi:10.1145/3643834.3660729.

27. P. Taylor, “Data Centers Worldwide by Country 2021,” *Statista*, Mar. 19, 2024. https://www.statista.com/statistics/1228433/data-centers-worldwide-by-country/

28. “Internet Users By Country 2021,” *worldpopulationreview.com*, 2021. https://worldpopulationreview.com/country-rankings/internet-users-by-country

29. D. A. Alsaleh, M. T. Elliott, F. Q. Fu, and R. Thakur, “Cross-cultural differences in the adoption of social media,” *JRIM*, vol. 13, no. 1, pp. 119–140, Mar. 2019, doi:10.1108/JRIM-10-2017-0092.

30. J. White *et al.*, “A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT,” Feb. 21, 2023, *arXiv*: arXiv:2302.11382. doi: 10.48550/arXiv.2302.11382.

31. L. Giray, “Prompt Engineering with ChatGPT: a Guide for Academic Writers,” *Annals of Biomedical Engineering*, vol. 51, pp. 2629–2633, Jun. 2023, doi: .org/10.1007/s10439-023-03272-4 .

32. S. Schulhoff, “Shot-Based Prompting: Zero-Shot, One-Shot, and Few-Shot Prompting,” *Learnprompting.org*, 2024. https://learnprompting.org/docs/basics/few_shot?srsltid=AfmBOorxtHkTVSCIDePh0lVBRhQtNQ6--HWP4oFkOdpzj3nNcZnwhslW (accessed Mar. 06, 2025).

33. “Cluster analysis,” *www.statskingdom.com*. https://www.statskingdom.com/cluster-analysis.html

34. Y. Song, G. Wang, S. Li, and B. Y. Lin, “The Good, The Bad, and The Greedy: Evaluation of LLMs Should Not Ignore Non-Determinism,” Jul. 15, 2024, *arXiv*: arXiv:2407.10457. doi: 10.48550/arXiv.2407.10457.

35. “OpenAI Platform,” *Openai.com*, 2025. https://platform.openai.com/docs/advanced-usage

36. L. Salewski, S. Alaniz, I. Rio-Torto, E. Schulz, and Z. Akata, “In-Context Impersonation Reveals Large Language Models’ Strengths and Biases,” Nov. 26, 2023, *arXiv*: arXiv:2305.14930. doi: 10.48550/arXiv.2305.14930.